ChatGPT: The End of Online Exam Integrity?

Abstract

This study addresses the significant challenge posed by the use of Large Language Models (LLMs) such as ChatGPT on the integrity of online examinations, focusing on how these models can undermine academic honesty by demonstrating their latent and advanced reasoning capabilities. An iterative self-reflective strategy was developed for invoking critical thinking and higher-order reasoning in LLMs when responding to complex multimodal exam questions involving both visual and textual data. The proposed strategy was demonstrated and evaluated on real exam questions by subject experts and the performance of ChatGPT (GPT-4) with vision was estimated on an additional dataset of 600 text descriptions of multimodal exam questions. The results indicate that the proposed self-reflective strategy can invoke latent multi-hop reasoning capabilities within LLMs, effectively steering them towards correct answers by integrating critical thinking from each modality into the final response. Meanwhile, ChatGPT demonstrated considerable proficiency in being able to answer multimodal exam questions across 12 subjects. These findings challenge prior assertions about the limitations of LLMs in multimodal reasoning and emphasise the need for robust online exam security measures such as advanced proctoring systems and more sophisticated multimodal exam questions to mitigate potential academic misconduct enabled by AI technologies.